A NONPREJUDICIAL METHOD OF DEFINING CLASS INTERVALS FOR COMPARING FREQUENCY DISTRIBUTION BY CHI-SQUARE

A Query by Hart Fischer, Dept. of Zool., Univ. of Toronto

D. S. Robson

BU-366-M April, 1971

Abstract

The comparison of several frequency distributions by means of a contingency chi-square test can be inadvertently biased through the selection of class intervals. Interval specification guided only by the combined array of row and column totals, however, cannot prejudice the outcome of the test. A nonprejudicial rule which results in approximately equal class frequencies for the combined samples is given by the two constraints

$$\text{number of classes} = c \leq \frac{\text{smallest sample size}}{5}$$

$$\text{total frequency in each class} \geq \frac{\text{total sample size}}{c} \quad .$$

A NONPREJUDICIAL METHOD OF DEFINING CLASS INTERVALS FOR COMPARING FREQUENCY DISTRIBUTIONS BY CHI-SQUARE

A Query by Hart Fischer, Dept. of Zool., Univ. of Toronto

D. S. Robson

BU-366-M April, 1971

Without meaningful guidelines for defining class intervals when two or more frequency distributions are to be compared by a chi-square test the statistical amateur may inadvertently bias the outcome of the test. This possibility is eliminated by the simple device of combining the several samples into a single sample before considering the question of class intervals; interval construction guided only by the combined array cannot prejudice the outcome of the test.

Since the sampling distribution of the contingency chi-square statistic is derived conditionally for fixed row totals (fixed sample sizes) and fixed column totals (fixed class frequencies for the combined sample) then a valid rule for determining class intervals is the a priori specification of both the number of classes c and the total class frequencies $N_1, N_2, \ldots, N_c$ (which must add up to the total combined sample size). In the case of continuous variables where ties cannot occur, any prior specification of class frequencies $N_1, N_2, \ldots, N_c$ summing to the combined sample size is achievable by ordering the observations in the combined sample by numerical value and assigning the first N_1 values to class 1, the next N_2 values to class 2, and so on. With discrete data where ties do occur, any given prior specification of class totals may be achievable only through the device of randomly splitting ties; a more practical procedure in this case is to choose a specification which does not require the splitting of ties. So long as this specification is based only upon inspection of the combined sample, ignoring the identification with individual samples, then the resulting chi-square test remains valid.

A reasonable though not necessarily optimal rule for specifying c and $N_1, N_2, \ldots, N_c$ is

$$c \leq \frac{\text{smallest sample size}}{5}$$

and

$$N_i \geq \frac{\text{combined sample size}}{c} .$$

Thus, this rule specifies approximately equal class frequencies for the combined sample, and the number of classes is chosen so that no expected cell frequency is less than 5. Applying this rule to the following discrete example gives

	0	1	2	3	4	5	6	7	Sample Size
Sample 1	27	15	8	5	5	6	2	1	69
Sample 2	48	31	17	8	6	8	3	3	124
Sample 3	124	80	50	20	17	7	2	8	308
Combined Sample	199	126	75	33	28	21	7	12	501
Class	199	126	75	61			40		

Here we have

$$c \leq \frac{69}{5} \doteq 14$$

and

$$N_i \geq \frac{501}{69/5} \doteq 36$$

Since 199 > 36 then $N_1 = 199$, and since 126 > 36 then $N_2 = 126$, and so on, resulting in an actual c of c = 5.